

DOI:10.3969/j.issn.1001-4551.2021.02.009

# 轴承尺寸检测数据的异常值 检测与数据处理研究 \*

何高清, 肖 健

(合肥工业大学 机械工程学院, 安徽 合肥 230009)

**摘要:**针对轴承尺寸检测系统中检测数据存在异常值和数据波动的问题,对异常值检测的方法和轴承尺寸检测数据的分布特点进行了研究,对异常值的产生与数据波动的原因进行了归纳,提出了一种基于箱型图理论异常值检测与最小二乘多项式拟合相结合的方法。首先,运用箱型图理论对检测数据进行了异常值筛选,再使用检测数据的中位数暂代了异常值;然后,利用最小二乘多项式拟合法对异常值暂代处进行了校正,且通过拟合的方式重新估计了检测数据;最后,通过轴承检测设备对其进行了实验验证。实验结果表明:该方法可以快速、高效地识别轴承尺寸检测数据中的异常值,有效地降低检测数据的波动性,箱型图法异常值识别率为 7.5%,高于  $3\sigma$  准则法 2.3%;最小二乘多项式拟合法可降低检测数据 50% 的波动性,显著提高检测结果的准确性。

**关键词:**轴承尺寸检测;异常值;箱型图;中位数;最小二乘多项式拟合

中图分类号:TH133.3;TP274

文献标识码:A

文章编号:1001-4551(2021)02-0198-06

## Outlier detection and data processing of bearing dimension detection data

HE Gao-qing, XIAO Jian

(School of Mechanical Engineering, Hefei University of Technology, Hefei 23009, China)

**Abstract:** Aiming at the problems of the existence of outliers and data fluctuations in bearing dimension detection system, the method of outliers detection and the distribution characteristics of data of bearing dimension detection were analyzed, the causes of outliers and data fluctuation were summarized, a method based on boxplot theory combining outliers detection with least square polynomial fitting was proposed. Firstly, the boxplot theory was used to filtrate the outliers. Then the median of the detection data was used to substitute outliers. The least square polynomial fitting method was used to correct the abnormal data, and the detection data was reassessed by this method. Finally, the test was verified by bearing detection equipment. The experimental results indicate that this method can identify the outliers in the bearing dimension detection data quickly and efficiently, and reduce the fluctuation of the detection data effectively. The outlier recognition rate of boxplot method is 7.5%, 2.3% higher than that of  $3\sigma$  method. The least square polynomial fitting method can reduce the fluctuations of detection data by 50% and improve the accuracy of detection results significantly.

**Key words:** bearing dimension detection; outliers; boxplot; median; least square polynomial fit

## 0 引言

在轴承形廓质量检测中,尺寸检测是重要的检测环节。本文采用多激光传感器并行高速自动化轴承检测设备,在连续尺寸检测过程中,其采样检测点数量众

多,检测数据存在异常值与数据波动是不可避免的。正确地识别与处理异常值、降低数据的波动性,对轴承检测结果的准确性与稳定性有重要意义。检测数据发生波动与产生异常值有以下几个主要原因:

(1) 由于选用激光传感器进行检测,其检测灵敏

收稿日期:2020-06-03

基金项目:合肥市远大轴承锻造有限公司资助项目(W2016JSKF0039)

作者简介:何高清(1977-),男,安徽合肥人,副教授,硕士生导师,主要从事数控技术与数据分析方面的研究。E-mail:hegq2008@163.com

度高,外界易对其造成干扰,造成检测数据不稳定;

(2)由于采用类三爪卡盘式固定轴承,当检测到卡盘处时,超出传感器检测量程,会产生异常数据点;

(3)由伺服电机带动卡盘高速旋转,进而对轴承进行检测,高速旋转所带来的振动也会影响检测结果。

异常检测也叫异常挖掘,是指从大量数据中找出其行为明显不同于预期对象的过程<sup>[1]</sup>。目前,异常数据检测的方法大体可分为基于统计的异常检测方法<sup>[2]</sup>、基于距离的异常检测方法<sup>[3]</sup>、基于密度的异常检测方法<sup>[4]</sup>和基于聚类的异常检测方法<sup>[5]</sup>等几种。各方法的优缺点分述如下:

(1)基于统计的异常检测方法,通过统计学理论,确定数据的分布模型,分析数据的离散程度和相应模型的评价指标来确定数据的异常程度,这种方法用于分析只包含单种属性的数据;

(2)基于距离的异常检测方法,通过设定距离阈值,计算各数据点与数据集的距离,将大于距离阈值的数据确定为异常数据。该方法不需要数据的具体分布模型,但其算法复杂度较高,不适用于大数据集和密度不均匀的数据集;

(3)基于密度的异常检测方法,能够检测出基于距离异常算法不能识别的一类数据——局部异常,打破了固有的绝对异常的观点,更符合实际应用,但其结果对参数的选择敏感,异常因子阈值的选取需要一定的先验知识;

(4)基于聚类的异常检测方法,一般利用 K-Means 算法将整个数据集聚类成多个簇,根据假设(异常点不属于任何的簇、异常点一般离最近的簇较远、稀疏簇中的点都被认为是异常的)确定异常数据,但其分类结果依赖于分类中心的初始化,对类别规模差异太明显数据的处理效果不好<sup>[6-9]</sup>。

本文中检测数据仅关于轴承尺寸属性,而且其数据量大,对系统实时性要求高,因而需要降低检测算法时间复杂度;综合以上异常检测方法的优势与不足并结合轴承尺寸检测数据的分布特点,笔者采用统计学箱型图理论对异常值进行检测,对于异常值用中位数暂代,再利用最小二乘多项式拟合法对原数据异常点处进行校正,且通过该方法对检测数据重新估计,提高检测结果的精度。

## 1 轴承尺寸检测系统

轴承检测系统主要由硬件系统与软件系统组成,设备实物图如图 1 所示。

其硬件系统主要包括:激光传感器、光栅尺、HMI

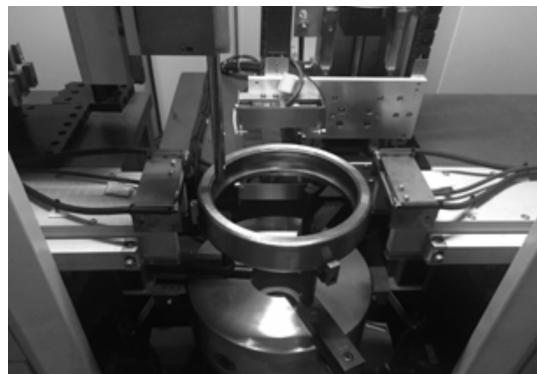


图 1 轴承尺寸检测设备

触摸屏、伺服驱动器及电机等。

其中,检测系统的硬件:激光传感器选用德国米铱 1420 型号,检测精度 1 μm,完成对轴承的尺寸采样;光栅尺用以记录传感器的位置,并将位置信号送入到 DSP 中;HMI 触摸屏,用以控制整个检测设备,并对不同种类轴承选择相应的测量方案;伺服驱动器及电机,用以将激光传感器移动到检测位置,并带动轴承的旋转运动。

软件系统主要包括 DSP(数字信号处理器),完成对电机的控制、与 HMI 触摸屏的信息交互、检测参数的处理、检测结果的输出等工作。

轴承产品的合格与否,根据检测结果的最大值、最小值是否在轴承的极限尺寸范围之内判断。

## 2 异常值的检测与数据处理

### 2.1 箱型图理论

基于正态分布的  $3\sigma$  准则是以假定数据服从正态分布为前提的,但实际数据往往并不符合正态分布模型,其以均值和方差为基础来判定数据的异常,受异常值本身的影响较大。而箱型图理论无需对数据做出限制,不受异常值的影响,可以直观地描述数据的离散分布情况,并且提供了一个识别异常值的标准,即大于箱型图设定的上界或小于下界的数值即为异常值。

箱型图如图 2 所示。

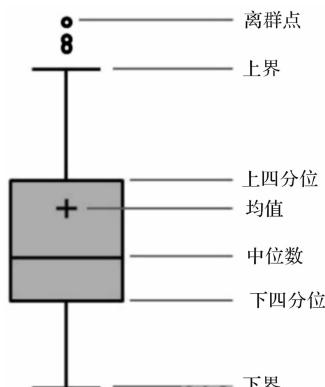


图 2 箱型图

将检测数据按照从小到大的顺序依次排列  $X_1, X_2, \dots, X_n$ , 得到有序数列, 则其中位数  $M$  记为:

$$M = \begin{cases} X_{\frac{n+1}{2}} & , n \text{ 为奇数} \\ \frac{1}{2} \left( \frac{X_{\frac{n}{2}}}{2} + X_{(\frac{n}{2}+1)} \right) & , n \text{ 为偶数} \end{cases} \quad (1)$$

异常值的判定标准为:

$$X_i > U + K \cdot IQR \text{ 或 } X_i < L - K \cdot IQR \quad (2)$$

式中:  $U$ —上四分位数, 区间  $[M, X_n]$  的中位数, 表明样本中只有  $1/4$  的数值大于  $U$ ;  $L$ —下四分位数, 区间  $[X_1, M]$  的中位数, 表明样本中只有  $1/4$  的数值小于  $L$ ;  $IQR$ —四分位距,  $IQR = U - L$ ;  $K$ —步长系数, 取  $K = 1.5$ 。

## 2.2 中位数暂代

选取中位数暂代异常值, 数据集的中位数比平均值具有更强的鲁棒性, 理论上可以“容忍”不超过总数据量  $50\%$  的异常值<sup>[10]</sup>, 并且保证了数据的完整性, 有助于对整体数据的最小二乘多项式拟合, 即:

$$X_i = M(X_i > U + K \cdot IQR \text{ 或 } X_i < L - K \cdot IQR) \quad (3)$$

## 2.3 最小二乘法多项式拟合

原始检测数据经过异常值检测与替代, 继而需要对替换值进行校正。使用最小二乘法多项式拟合的方式, 可以根据整体数据的分布趋势对替换值进行校正。同时, 利用拟合的方式重新处理数据, 降低其波动性。

### 2.3.1 最小二乘法基本原理

最小二乘法(最小平方法)是一种数学优化技术, 通过最小化误差的平方和寻找数据的最佳函数匹配<sup>[11]</sup>。

对于一组实验数据  $(x_i, y_i) (i = 0, 1, \dots, m)$ , 要求在某个函数类,  $\Phi = \text{span}\{\varphi_0(x), \varphi_1(x), \dots, \varphi_n(x)\}$  中寻求一个函数, 即:

$$\begin{aligned} \varphi(x) &= \alpha_0 \varphi_0(x) + \alpha_1 \varphi_1(x) + \dots + \\ \alpha_n \varphi_n(x) &= \sum_{k=0}^n \alpha_k \varphi_k(x) \end{aligned} \quad (4)$$

使  $\varphi(x)$  满足条件:

$$\begin{aligned} \sigma &= \min \sum_{i=0}^m [\varphi(x_i) - y_i]^2 = \\ \min \sum_{i=0}^m &[ \sum_{k=0}^n \alpha_k \varphi_k(x_i) - y_i ]^2 \end{aligned} \quad (5)$$

由多元函数极值必要条件可知:

$$\frac{\partial \sigma}{\partial \alpha_j} = 2 \sum_{i=0}^m [ \sum_{k=0}^n \alpha_k \varphi_k(x_i) - y_i ] \varphi_j(x_i) = 0 \quad (0 < j < n) \quad (6)$$

即:

$$\frac{\partial \sigma}{\partial \alpha_j} = 2 \sum_{k=0}^n [ \sum_{i=0}^m \varphi_j(x_i) \varphi_k(x_i) ] \alpha_k = \sum_{i=0}^m y_i \varphi_j(x_i) \quad (7)$$

记:

$$(h, g) = \sum_{i=0}^m h(x_i) g(x_i) \quad (8)$$

则上式可表示为:

$$\alpha_0 (\varphi_j, \varphi_0) + \alpha_1 (\varphi_j, \varphi_1) + \dots + \alpha_n (\varphi_j, \varphi_n) = (\varphi_j, y) \quad (9)$$

写成矩阵形式为:

$$\begin{bmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) & \dots & (\varphi_0, \varphi_n) \\ (\varphi_1, \varphi_0) & (\varphi_1, \varphi_1) & \dots & (\varphi_1, \varphi_n) \\ \vdots & \vdots & \ddots & \vdots \\ (\varphi_n, \varphi_0) & (\varphi_n, \varphi_1) & \dots & (\varphi_n, \varphi_n) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} (\varphi_1, y) \\ (\varphi_2, y) \\ \vdots \\ (\varphi_n, y) \end{bmatrix} \quad (10)$$

式(10)即为最小二乘法求解的法方程组。

根据方程组求解得  $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_n)^T$ , 则  $\varphi(x) = \sum_{k=0}^n \alpha_k \varphi_k(x)$  为所求拟合函数。

### 2.3.2 拟合函数的选取

合适的  $\varphi(x)$  可以增强模型对检测数据的解释能力。

以轴承内径尺寸检测为例, 本文选用拟合函数为多项式函数, 是根据在 MATLAB 的 Curve Fitting 工具箱中<sup>[12]</sup>, 利用有理函数、三角函数和多项式函数对筛选和替换后的检测数据拟合后所得到的。

各拟合函数效果如图 3 所示。

对于曲线拟合效果是否最佳, MATLAB 有具体的评价指标 SSE 和 R-square。其中, SSE 为误差平方和, 该参数计算拟合参数后的回归值与原始数据对应点的误差平方和, SSE 越小说明模型选择和拟合得更好; R-square 为确定系数, 其值越接近 1, 表明方程的自变量对因变量的解释能力越强, 模型对数据的拟合程度越好。

各曲线拟合程度评价指标如表 1 所示。

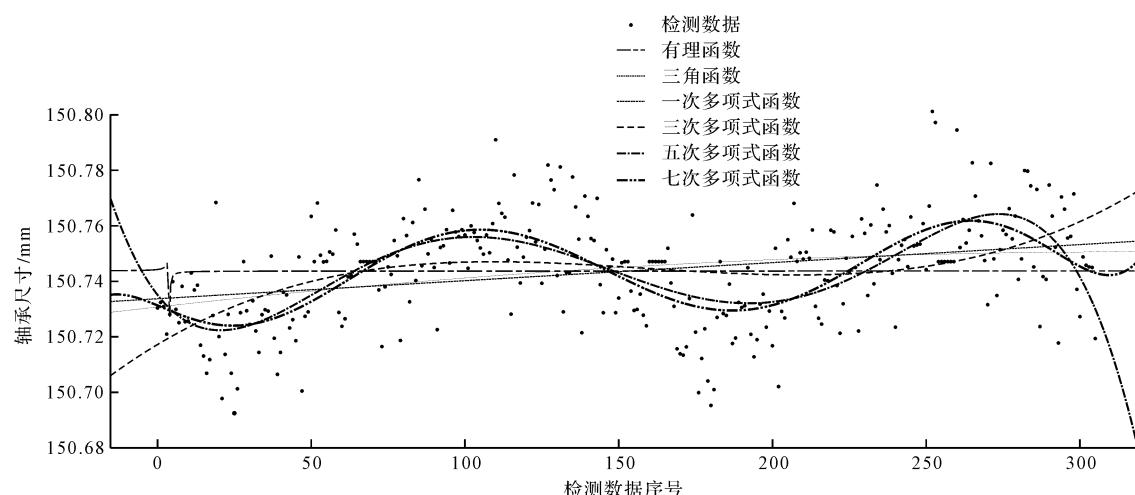


图3 各拟合函数效果

表1 各曲线拟合程度评价指标

函数模型	SSE	R-Square
三角函数	0.099 82	0.102 5
有理函数	0.097 86	0.120 1
一次函数	0.100 3	0.098 46
三次函数	0.116 3	0.147 1
五次函数	0.067 01	0.397 5
七次函数	0.063 61	0.428

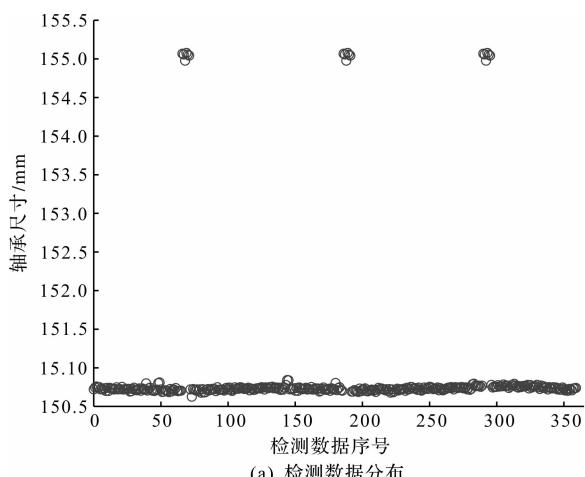
根据表1中SSE与R-square综合考虑,笔者选择七次多项式函数为拟合函数,即:

$$\varphi(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 + \alpha_5 x^5 + \alpha_6 x^6 + \alpha_7 x^7 \quad (11)$$

### 3 轴承检测实验及结果分析

笔者以公称内径尺寸为Φ150.7 mm轴承检测为例,进行实验。

检测数据分布与频率分布直方图,如图4所示。



(a) 检测数据分布

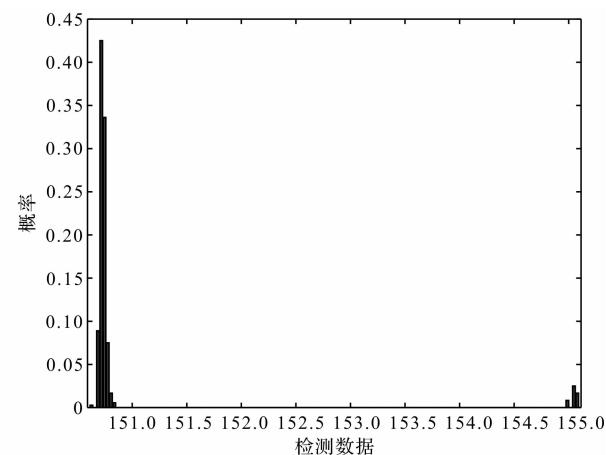
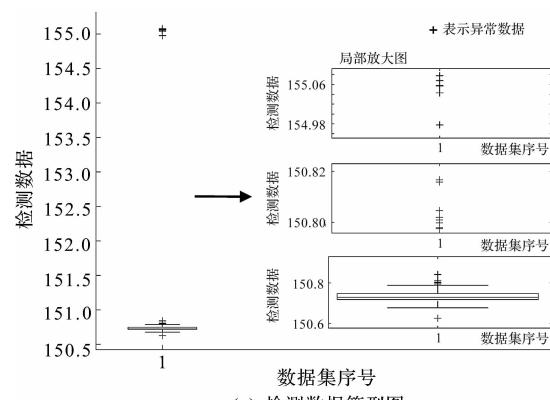


图4 检测数据分布与频率分布直方图

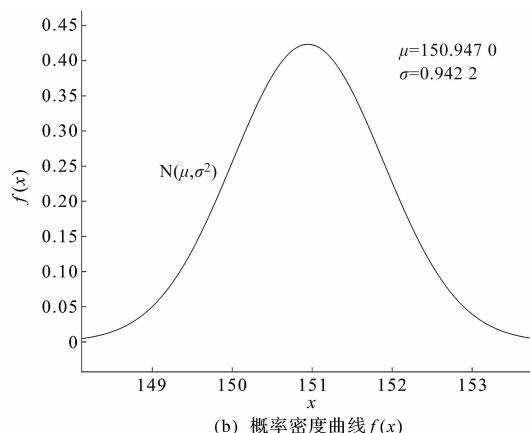
根据图4,通过计算可得到检测数据的均值  $\mu = 150.947 0$  和标准差  $\sigma = 0.942 2$ ,则正态分布的概率密度曲线  $f(x)$  为:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (12)$$

检测数据箱型图和概率密度曲线  $f(x)$  如图5所示。

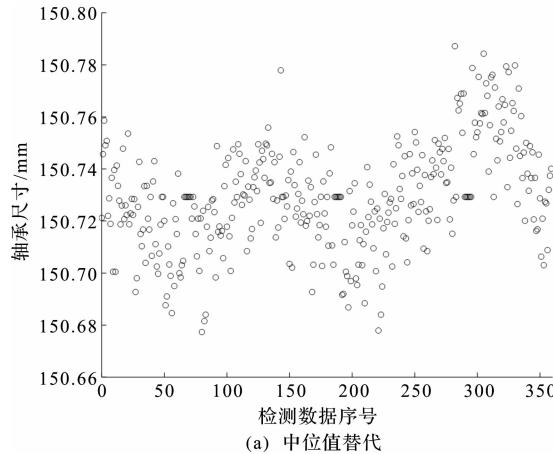


(a) 检测数据箱型图

图 5 检测数据箱型图和概率密度曲线  $f(x)$ 

当传感器检测到卡盘处时,会得到如图 4(a)中的上部异常数据。由图 5(a)观察可知,箱型图可以检测出这类异常数据以及其他原因所造成的异常值。由图 4(b)与图 5(b)对比可知,检测数据的实际分布模型不符合正态分布。

笔者分别使用箱型图法与  $3\sigma$  准则法识别数据的



(a) 中位数替代

异常值,异常数据检测结果如表 2 所示。

表 2 异常数据检测结果

检测方法	箱型图法	$3\sigma$ 准则法
检测数据个数	360	360
检测数据均值	150.947 0	150.729 2
检测数据中位数	150.792 3	153.773 7
识别异常值个数	27	19
异常值识别率	7.5%	5.2%
异常值识别上限	150.673 1	148.120 3
异常值识别下限	150.673 1	148.120 3

由表 2 结果对比可知:箱型图法的异常值识别率高于  $3\sigma$  准则法 2.3%,而且箱型图法的异常值识别区间小于  $3\sigma$  准则法,表明箱型图法对异常值的识别准确率更高;主要由于  $3\sigma$  准则法要求数据服从正态分布,然而实际数据分布并不能满足要求。

因此,在大数据量的检测系统中,箱型图法更具有优势,故在本研究中选用箱型法作为检测异常值的方法。

笔者分别用检测数据的中位数与平均值替换异常值。异常值替换后数据分布如图 6 所示。

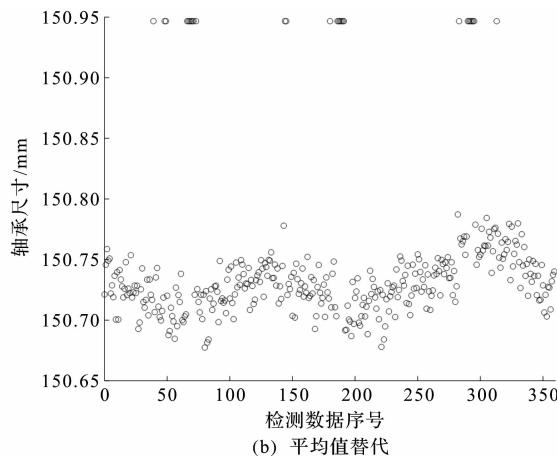


图 6 异常值替换后数据分布

由图 6 可知:用中位数替代异常值数据分布更为集中化,均值由于受异常值影响较大,而且其在箱型图法中属于异常值,不适合选作替换值。

综上所述,笔者认为选取中位数更为合适。

笔者对用中位数替换后的数据进行拟合。七阶多项式最小二乘拟合结果如图 7 所示。

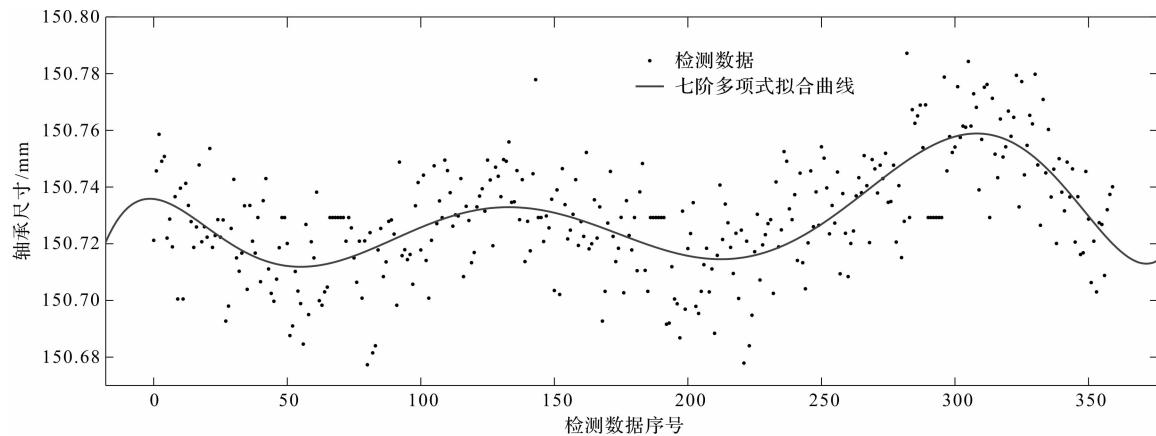


图 7 七阶多项式最小二乘拟合

由图7可知,未经拟合时数据的波动范围约为0.1 mm,七阶多项式拟合后数据的波动范围约为0.05 mm,明显地降低了数据的波动性;拟合的数据分布相比较直接用中位数替代的数据分布,更符合实际测量趋势。

## 4 结束语

针对轴承检测系统中出现的异常值与数据波动问题,笔者采用统计学箱型图理论对异常值进行检测,对于异常值用中位数暂代,再利用最小二乘多项式拟合法对原数据异常点处进行校正,且通过该方法对检测数据重新估计。

实验及研究结果表明:

- (1) 箱型图法异常值识别率高于 $3\sigma$ 准则法2.3%,可准确、快速地识别异常值;
- (2) 中位数替换的方式受异常值影响较低,保证了检测数据的完整性;
- (3) 通过最小二乘多项式拟合的方式,数据波动降低为原来的50%,使数据分布更为合理化。

此方法的时间和空间复杂度低,易于实现编程,可保证检测系统的实时性和准确性。因此,对于使用位移传感器测量零件尺寸的系统具有一定的参考价值。

## 参考文献(References):

- [1] 孙建树,娄渊胜,陈裕俊. 基于ARIMA-SVR的水文时间序列异常值检测[J]. 计算机与数字工程,2018,46(2):225-230.
- [2] 戴邵武,陈强强,毛凯,等. 基于样本分位数原理的飞参数据异常值检测算法[J]. 兵器装备工程学报,2020,41(5):113-117.
- [3] 聂志红,阚常壮,谢扬. 连续压实质量检测参数单点异常值识别及处理[J]. 哈尔滨工业大学学报,2019,51(3):150-157.
- [4] 董泽,贾昊. 基于EWT-LOF的热工过程数据异常值检测方法[J]. 仪器仪表学报,2020,41(2):126-134.
- [5] 陆春光,叶方彬,赵羚,等. 基于密度峰值聚类的电力大数据异常值检测算法[J]. 科学技术与工程,2020,20(2):654-658.
- [6] 何奇峰. 考虑季节性和趋势性影响的时空数据异常值检测研究[D]. 北京:北京邮电大学管理科学与工程学院,2018.
- [7] GUO Z M. Fast outlier detection algorithm based on local density and connectivity[J]. *Scientific Journal of Intelligent Systems Research*, 2020, 2(1):18-27.
- [8] GAN G J, NG M K. K-means clustering with outlier removal [J]. *Pattern Recognition Letters*, 2017(90):8-14.
- [9] 杨晓玲,冯山,袁钟. 基于相对距离的反k近邻树离群点检测[J]. 电子学报,2020,48(5):937-945.
- [10] 刘峻清,陶涛. 一种污水处理RO膜压差异常数据检测和处理方法[J]. 四川环境,2019,38(1):13-17.
- [11] 卢治功,贺鹏,职连杰,等. 基于最小二乘法多项式拟合三角测量模型研究[J]. 应用光学,2019,40(5):853-858.
- [12] 刘利敏,吴敏丽. 基于MATLAB的最小二乘曲线拟合[J]. 福建电脑,2019,35(8):9-12

[编辑:雷敏]

## 本文引用格式:

何高清,肖健. 轴承尺寸检测数据的异常值检测与数据处理研究[J]. 机电工程,2021,38(2):198-203.

HE Gao-qing, XIAO Jian. Outlier detection and data processing of bearing dimension detection data[J]. Journal of Mechanical & Electrical Engineering, 2021, 38(2):198-203.

《机电工程》杂志: <http://www.meem.com.cn>